A
Synopsis entitled
“Real-Time Multilingual Sentiment Analysis On Social Media Data Using Transformer-Based Natural Language Models”

Submitted to

Swami Ramanand Teerth Marathwada University,
Vishnupuri, Nanded

in partial fulfillment for the Registration of the degree, Doctor of Philosophy in Computer Science (Faculty of Science & Technology)

Submitted By

Nahar Sunny Suresh
School of Technology, S.R.T.M. University Nanded, Sub-Center, Peth, Latur

Under the guidance of

Dr. Prashant Prakashrao Agnihotri
Associate Professor,
School of Technology,
Swami Ramanand Teerth Marathwada
University, Sub-Centre, Peth, Latur , India.

December 2024

Contents

1. Abstract 3
2. Introduction 3
3. Motivation and Scope 5
4. Literature of Review 5
   4.1. Research gaps 8
5. Problem Statement 8
6. Research Objectives 8
7. Research Methodology 9
   7.1. Dataset Description 9
   7.2. Techniques Used 9
   7.3. Proposed Methodology 12
8. Expected Outcome 14
9. Research Work Plan 15
10. References 15

Real-Time Multilingual Sentiment Analysis On Social Media Data Using Transformer-Based Natural Language Models

1. Abstract

The rapid growth of multilingual data in digital platforms has sparked the need for real-time sentiment analysis that can address linguistic diversity and computational scalability. This study investigates the improvement of multilingual Natural Language Processing (NLP) models for efficient distributed sentiment analysis. The proposed method combines innovation in preprocessing techniques: image text preprocessing using Optical Character Recognition (OCR), feature engineering by Term Frequency-Inverse Document Frequency (TF-IDF), and dimensionality reduction by Principal Component Analysis (PCA), all while utilizing advanced architectures based on transformers. The proposed framework leverages a robust methodology in training and evaluating the models with deep learning and transformer models in multilingual datasets so that it can handle code-mixed languages, limited resource languages, and real-time applicability. Continuous performance monitoring would enable adaptation to dynamic multilingual streams. Scalable high-accuracy sentiment analysis could be one of the expected outcomes for the marketing, public policy, and customer feedback analysis industries. This study fills in the significant gaps found in multilingual sentiment analysis, which involves efficiency improvement, integration, and better accuracy toward the detection of sentiments, ultimately advancing the deployment of distributed real-time NLP-based systems.

Keywords: Multilingual Sentiment Analysis, emotion detection, social media, deep learning, transformers, Natural Language Processing (NLP).

2. Introduction

The exponential growth of digital communication brought individuals and organizations closer to communicating worldwide, producing vast amounts of multilingual text data across several social media services, e-commerce platforms, and other digital channels [1]. Sentiment analysis emerges as a key domain for Natural Language Processing (NLP) to get insights into public opinion, customer satisfaction, and trends [2]. However, the multifaceted linguistic nature of the user-generated content poses significant challenges in this regard, as conventional sentiment analysis models often fail with the subtleties and complexity of multiple languages [3,4]. Figure 1 illustrates the basic steps of sentiment analysis.

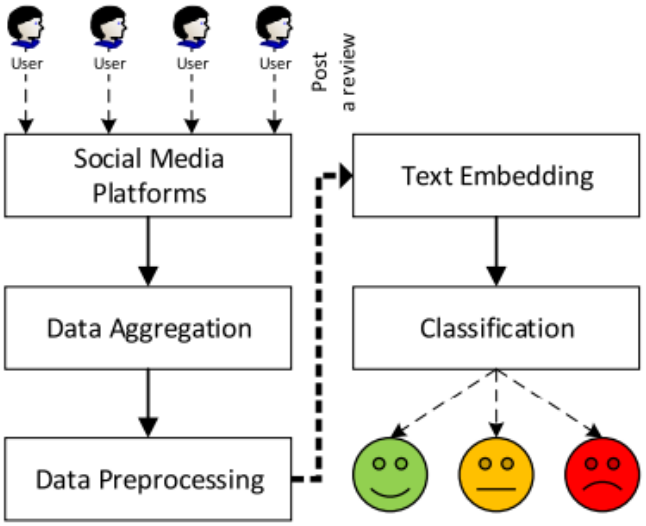

Figure 1: Basic steps of sentiment analysis [5].

In this context, it has become vital to perform real-time multidimensional distributed sentiment analysis for NLP models in multiple languages [6]. The marketing, political, customer-oriented or healthcare businesses need to process sentiments in a multilingual context in real-time. Real-time sentiment insights are very important in tactical decision-making processes [7,8]. These advances in machine learning, deep learning, and distributed computing make it possible for states to address the problems of linguistic and computational multiculturalism while analyzing the streams of multilingual data [9,10].

The transformer-based architectures and their multilingual variants that have proven to be effective in text comprehension and generation across languages are an essential element of this improvement [11,12]. Such models incorporate distributed computing paradigms like Apache Kafka and Spark, and therefore, can handle big data in real time, which makes them scalable and efficient [13,14]. In addition, the integration of transfer learning, target domain learning, or tuning on various datasets enhances language-specific sentiment understanding [15].

Such advancements also appeal to broader uses apart from sentiment analysis. Improved multilingualism enhances coverage, allowing businesses and policymakers to effectively engage with different groups of people [16]. Such models also enable worldwide surveys of these trends and opinions, enhancing competitiveness in the knowledge of market trends [17]. However, achieving these goals comes with significant challenges, including ensuring computational efficiency, addressing privacy concerns highlighted by the Data Protection Act, and mitigating biases inherent in the training datasets [18].

In essence, Real-Time Multilingual Sentiment Analysis On Social Media Data Using Transformer-Based Natural Language Models is a significant leap in the field of NLP and has

great potential for modifying how emotions are constructed and understood at a global level and in a multilingual context. Placing a kind of focus on scale, precision, and inclusivity brings these systems closer to the possibility of operating as efficient means of evaluating and responding to worldwide sentiments instantly.

3. **Motivation and Scope**

The growing amount of multilingual content being created all across the world is evidence of the demand for complex systems that allow sentiment analysis to be executed in real-time. Many traditional approaches do not solve the issues of linguistic diversity and the need for low-cost, high-speed consecutive processing, thus leaving a great void in the NLP domain. This research addresses these opportunities by improving multilingual NLP models for these tasks and at the same time adds value to the existing body of knowledge by proposing strategies that are cognizant of the rich language and the low resource computational requirements. The importance of this research lies in its capability to deliver quick responses to the many diverse comments from the public enabling the affected industries such as marketing, public policy or social monitoring to make better decisions. Also, it bolsters decentralized Artificial Intelligence (AI) systems through the fusion of multilingual language models with multi-tenant frameworks. Not only does this study fill the existing voids but it also paves the way for subsequent breakthroughs which add more robust frameworks and methods for global sentiment analysis to the field.

4. **Literature of Review**

This section covers a thorough overview of the existing research conducted by different authors addressing the Real-Time Multilingual Sentiment Analysis On Social Media Data Using Transformer-Based Natural Language Models.

Liu et al., (2024) [19] investigated the use of various deep learning-based NLP technologies in sentiment analysis in languages other than English. This was because sentiment classification faced several challenges when dealing with multiple languages. The study employed Bidirectional Encoder Representations from Transformers (BERT) and Long-Short Term Memory (LSTM) models and assessed multilingual situations related to data processing capacity for various languages. Deep learning models were used and trained in multilingual contexts bilingual and trilingual. The models were shown to have achieved highly accurate results, with BERT reaching up to a 93% accuracy rate. The study emphasized the need for models to enhance their contextual comprehension, create effective context-dependent neural networks, and apply multilingual resource data to achieve quality and accuracy in cross-lingual sentiment analysis.

Hashmi et al., (2024) [20] identified and mitigated multi-lingual hate speech spanning across 13 languages to minimize its negative socio impact. The study implemented traditional

machine learning techniques, popular deep learning techniques, and transformer models enhanced by tuning and optimization strategies. Ethical and political issues were also dealt with using explainable AI. The models recorded significant improvements with F1-scores of 0.90 for the German and 0.95 for the Roman dataset. In the case of mixed language data sets, the accuracy scores improved from 0.59 to 0.96. The study described an effective and scalable architecture for the detection of hate speech that redefined the performance metric of detection across languages and datasets.

Sampath et al., (2024) [21] explored the machine translation challenges posed by code-mixed texts which had become a significant presence in Indian social media. The study fine-tuned IndicLID to use it for Language Identification (LID). Then, it employed the ensemble approach using IndicTrans and IndicXlit for transliteration and translation purposes, respectively. IndicLID achieved a higher accuracy of more than 99% for language identification, whereas code-mixed data showed more than 95% accuracy. Traditional approaches to machine translation were significantly outperformed by the proposed methods. The approach properly translated code-mixed texts, thus improving sentiment analysis for Hinglish and all other Indian language combinations.

Ghosh et al., (2023) [22] aimed at improving sentiment detection on code-mixed data through emotion recognition in a multitask learning setup. The study proposed a multitask system based on XLMR, which used cross-lingual embedding-based transformer models and transfer learning. The model achieved an accuracy of 71.61% on the sentiment task, outperforming the best available systems by 1.93%. On the emotion task, the accuracy was 63.47%, which showed better generalization and semantic coverage. Multiple related objectives and transfer learning improved the current model performance and made it able to handle code-mixed Hinglish data.

Khan et al., (2023) [23] developed Election-focused BERT (ElecBERT) for the sentiment analysis of election-related tweets, addressing issues like sarcasm and misinformation. Built on BERT, ElecBERT was further fine-tuned on datasets such as the Election-Related Sentiment Annotated Tweets (ElecSent) dataset and the ElecSent-English dataset. It compared performance metrics against other models like Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost). On ElecSent-Multi-Languages, ElecBERT gained an accuracy of 99.05% and an F1-score of 98.16% against BERTweet and other baselines. The model achieved superior performance in the sentiment analysis of election-related tweets with applications in political analysis and social media management.

Yang et al., (2023) [24] addressed issues of disinformation and communication barriers that existed during the COVID-19 crisis by creating an AI chatbot called DR-COVID, that

processes the information in various languages. The multilingual chatbot incorporated an ensemble NLP model for nine languages and evaluated metrics such as Area Under Curve (AUC), precision, recall, and accuracy. DR-COVID achieved top 3 accuracies of 92.2% with an AUC score of 96.0%, best results for the Portuguese language as non-English 90%. DR-COVID demonstrated significant potential for pandemic education and multilingual healthcare delivery.

Kanfoud et al., (2022) [25] proposed SentiCode, which was a language-independent sentiment analysis approach without relying on any machine translation. SentiCode used the part-of-speech tags of Universal Dependencies to represent the languages in a non-dependent manner. It involved experiments with the machine and deep learning techniques. SentiCode demonstrated a real-time capability with an efficiency of 0.67 seconds during training and testing which in turn enabled it to yield an accuracy of 70%. The study also offered an effective and practical framework for multilingual sentiment analysis in different disciplines and languages.

Li et al., (2021) [26] focused on the development of a multi-language model of their emotion analysis to address the issues pertaining to the recognition of emotional polarization. The model integrated such multilingual documents by embracing emotional word embeddings that belong to the fast language model. The framework recorded an accuracy of 92% across the eight languages which proved its relevance and efficiency. The multilingual emotion analysis framework effectively addressed the existing knowledge deficiencies enabling the global intelligence on sentiment analysis.

El-Affendi et al., (2021) [27] suggested a new model the Multilevel Parallel Attention Neural (MPAN) model suited for Arabic Sentiment Analysis (ASA). The proposed model MPAN incorporated the use of positioning binary embedding schemes for contextualized embeddings. Results showed that the model was able to achieve better results than shallow machine learning and non-contextualized embedding schemes. The MPAN model enabled on greatly furthering of the ASA tasks by giving specific influence of the problems of interest in the ASA tasks on the model development.

Song et al., (2021) [28] focused on utilizing Twitter data to detect disasters which could facilitate emergency response planning. The study described the issues that arose from the fact that a Tweet was often very short and thus provided limited context and that many words in a disaster-related post were inherently vague. The proposed approach, SentiBERT-BiLSTM-CNN, integrated SentiBERT to generate sentiment contextual embeddings, Bidirectional LSTM (BiLSTM) with attention for contextual learning, and a 1D Convolutional Neural Network (CNN) layer for the extraction of local features. The model

achieved a remarkable F1 score of 92.75, significantly outperforming peer methods and proving its effectiveness for Tweets-based disaster prediction.

4.1. **Research gaps**

- Limited exploration of the role of cultural and linguistic diversity in improving cross-linguistic sentiment classification in multilingual NLP models [19].
- Lack of research on integrating advanced translation and sentiment analysis methods for real-time code-mixed text analysis [21].
- Inadequate focus on optimizing multilingual NLP models for real-time distributed sentiment analysis with low latency and high accuracy [25].
- Insufficient investigation into addressing underrepresented languages with limited labeled data in multilingual sentiment analysis frameworks [26].

5. **Problem Statement**

The increasing volume of multimedia data generated on social media and other digital platforms presents a significant challenge in analyzing sentiment across diverse languages and contexts in real-time. Existing sentiment analysis systems often struggle with limited scalability, inability to handle the variability and complexity of visual data, and lack of robustness in multilingual contexts. Traditional models are typically constrained by their reliance on textual data, neglecting the sentiment-rich visual content, and failing to generalize effectively across different sentiment categories such as positive, neutral, and negative. Furthermore, the imbalanced distribution of sentiment classes and the high-dimensional feature space of visual data exacerbate the inefficiency of current systems. These limitations highlight the need for advanced approaches that can integrate deep learning models, effectively preprocess visual data, and adapt to distributed environments for real-time multilingual sentiment analysis.

6. **Research Objectives**

The following are the major objectives of the study:

- To build and train sentiment classification models, including BERT, Robustly Optimized BERT Approach (RoBERTa), and LSTM, for analyzing sentiment categories.
- To develop an efficient framework for real-time sentiment classification using multilingual NLP models.

- To implement feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) and reduce feature dimensionality with Principal Component Analysis (PCA) for optimal performance in multilingual sentiment analysis.
- To identify areas for continuous improvement to enhance the robustness and adaptability of the sentiment analysis system in real-time scenarios.
- To evaluate the classification models using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to assess performance across multilingual sentiment classes.

7. **Research Methodology**

This study offers a robust analysis of the employed strategy, detailing the dataset, methods, and the recommendation approach as follows:

7.1. Dataset Description

The MVSA dataset [29] contains the image-text pairs that were retrieved from Twitter and the tweets have been made by more than three hundred million active users who create five hundred million Tweets every day. Employing the Twitter Application Programming Interface (API), Image-text pairs were collected from the tweets using a 406-keyword emotional vocabulary set that was limited and included 10 different emotions such as joy, love, and sadness. These keywords, organized into three degrees of emotional intensity (strong, medium, light), ensured a balanced and representative data collection by limiting each keyword to a maximum of 100 tweets per round. To diversify content, data collection occurred daily across multiple time slots. After filtering tweets containing emotional keywords or hashtags, image Uniform Resource Locator (URLs) embedded in the messages were extracted, and only tweets with accessible paired images and text were retained for annotation. This methodology ensures a rich and diverse dataset focused exclusively on visual sentiment representation.

7.2. Techniques Used

This section delineates the techniques utilized in the proposed methodology.

- **BERT**

BERT is a state-of-the-art language representation model designed to understand the context of words concerning surrounding text by employing a bidirectional transformer architecture [30]. Unlike traditional models that process text sequentially, BERT analyzes text in both forward and backward directions simultaneously, capturing deep contextual relationships [31]. It leverages pre-training on large corpora using tasks like Masked Language Modeling

(MLM) and Next Sentence Prediction (NSP), enabling it to perform effectively across diverse NLP tasks, such as sentiment analysis, question answering, and text classification [32]. BERT's multilingual variants, such as mBERT, extend its capabilities to multiple languages, making it highly suitable for complex multilingual sentiment analysis applications [33,34]. Figure 2 illustrates the architecture of the BERT model.

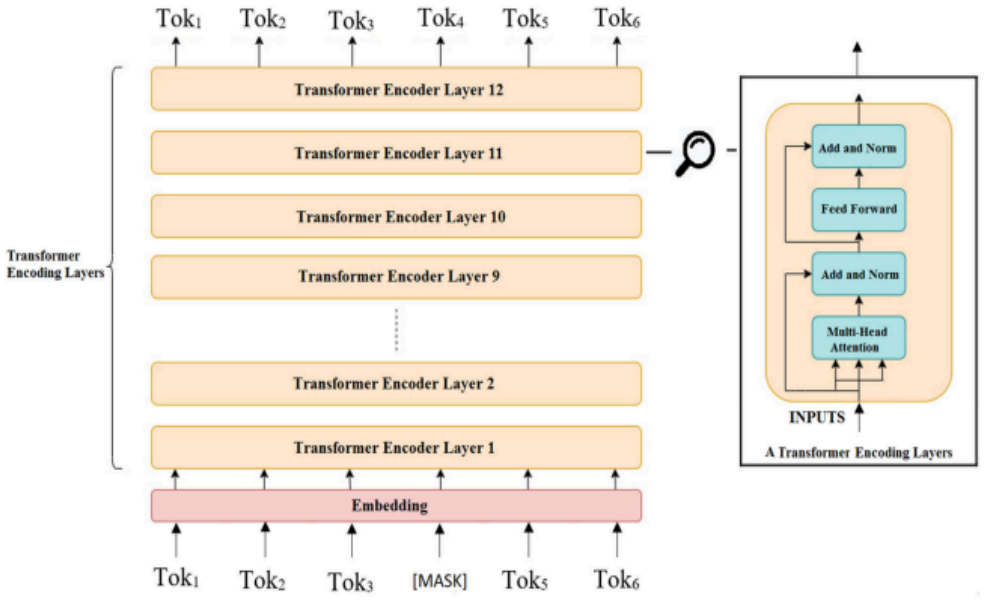

Figure 2: BERT model architecture [35]

- **Robustly Optimized BERT Approach (RoBERTa)**

RoBERTa's deviations from BERT come down principally to the optimization of the data utilized during pretraining and subsequent alterations made to the model, aimed at improving its functionality within the realm of natural language processing [36]. In contrast to BERT, RoBERTa omits the Next Sentence Prediction (NSP) objective, employs larger datasets when training, and trains for a larger number of epochs (and longer sequences) [37]. It combines the reinforcement of static notations with a finer structure of context depiction and employs various sources of data for pretraining including multilingual data sets, making it highly competitive in tasks such as question answering, language inference, and language sentiment across border uplift [38]. Such an improvement extends its ability to grasp the fine features associated with languages and makes it a suitable model for multilingual tasks in NLP [39].

- **LSTM**

LSTM is a specific type of Recurrent Neural Network (RNN) that can handle sequential data in the manner of overcoming vanishing gradient problems that are inherent to the traditional RNNs [40]. Memory cells of the LSTM networks use input, forget, and output gates to control information flow through the network to allow selective remembering or forgetting of

data along long sequences [41]. This feature makes LSTM highly suitable for applications that need contextual comprehensions, such as sentiment analysis, time-series forecasting, and language modeling, since it memorizes relevant features from previous inputs while eliminating the irrelevant ones [42]. The mechanism allows LSTM models to perform exceptionally well in detecting long-range dependencies and temporal patterns in multilingual sentiment analysis, thereby allowing the model to make accurate classification across different languages and types of sentiments [43]. Figure 3 illustrates the LSTM Architecture.

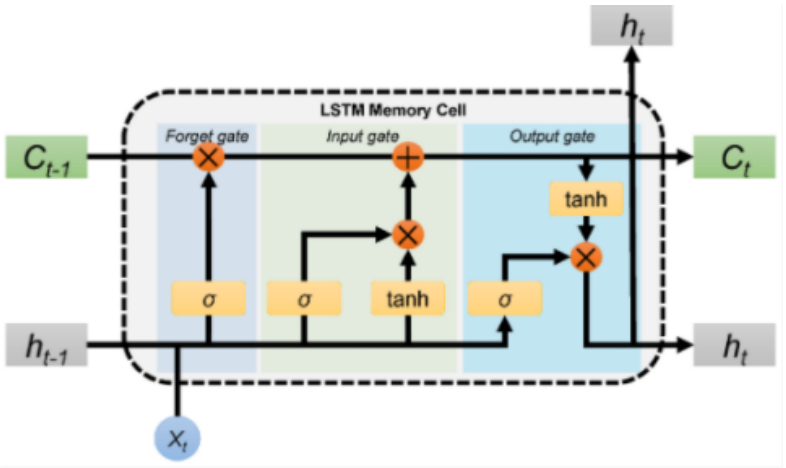

Figure 3: Architecture of LSTM [44]

- **Term Frequency-Inverse Document Frequency (TF-IDF)**

TF-IDF is a statistical method often used in NLP as well as information retrieval to assess the significance of a word inside a text in relation to a corpus of documents [45]. It integrates two metrics: TF, which quantifies the frequency of a phrase inside a document, indicating its significance in that context, and IDF, which diminishes the importance of terms prevalent throughout several documents, so highlighting words that are more distinctive to certain settings [46]. This value is then multiplied by another value called the TF-IDF score. This score has greater significance for terms appearing more often in a specific document but not throughout the corpus. This is the reason why TF-IDF can be efficiently applied as a feature extractor for tasks in texts, such as keyword extraction, sentiment analysis, and the assessment of similarity in documents [47].

- **Principal Component Analysis (PCA)**

PCA is one of the dimensionality reduction methods for transforming high-dimensional data while preserving maximum retention of variance into a lower dimensional space [48]. It is an approach that finds eigenvalues and eigenvectors decompositions of the covariance matrix such that the principal axes along orthogonal directions with maximum variation. PCA ranks these components in terms of explained variance so that one can choose a sub-selection that captures most of the information, thus reducing computational cost and the possibility of overfitting in machine learning applications [49]. Such a technique is highly useful in the

preprocessing of datasets to make data visualization easy and also improve models' efficiency without losing crucial information significantly [50]. Figure 4 depicts the visualization of the output from PCA.

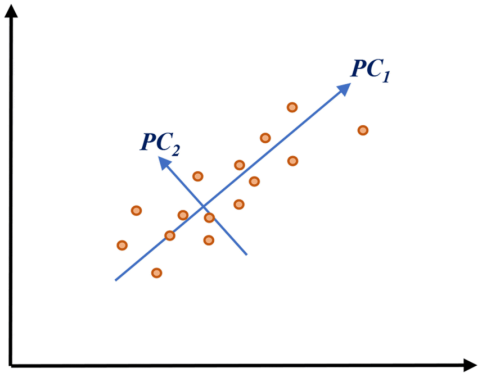

Figure 4: PCA [51]

- **Optical Character Recognition (OCR)**

OCR is a technology that enables the conversion of different types of documents, such as scanned paper documents, PDFs, or images captured by a camera, into editable and searchable data. It works by analyzing the shapes of letters and symbols in an image and translating them into machine-readable text [52]. In the context of this study, OCR plays a crucial role in extracting text from images associated with tweets in the MVSA dataset. Many tweets might contain images with embedded text, such as memes, screenshots, or text in images, which holds valuable sentiment-related information. Without OCR, this textual content would remain inaccessible for analysis, limiting the effectiveness of the sentiment analysis process.

7.3. Proposed Methodology

The methodology begins by using the MVSA Twitter dataset, which includes text and images, where text from images is extracted using OCR and merged with the tweet text. Data preprocessing involves removing stop words, tokenizing text, and applying lemmatization and stemming to standardize it across multiple languages. Features are then extracted using TF-IDF to emphasize important terms and reduced via PCA for dimensionality reduction. The dataset is split into training and testing subsets, ensuring multilingual balance. Three models: BERT, RoBERTa, and LSTM are trained to analyze sentiment, with possible ensemble approaches to enhance performance. The models are evaluated using metrics such as accuracy and F1-score, and the best-performing model(s) are deployed for real-time sentiment analysis. A continuous monitoring system ensures model performance is tracked, and periodic updates are made using new data to maintain accuracy and handle changes in multilingual text streams effectively. Figure 5 given below illustrates the proposed methodology in detail.

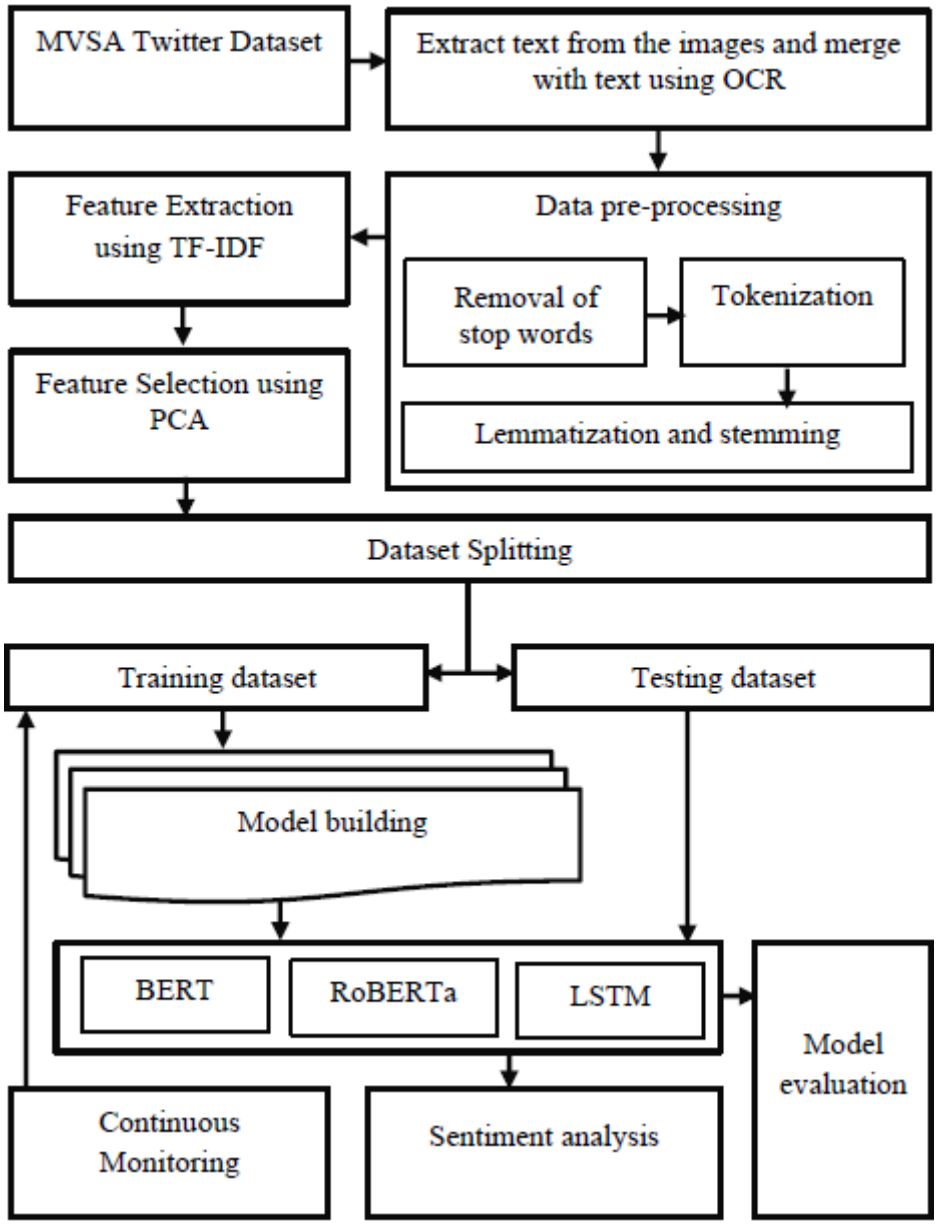

Figure 5: Proposed Methodology

7.3.1. MVSA Twitter Dataset

The dataset includes paired images and text from tweets, which are collected using emotional keywords. It provides a multilingual base for analyzing sentiments with visual and textual data.

7.3.2. Extract Text from Images

Optical Character Recognition (OCR) is used to extract text from images, combining it with tweet text for a more comprehensive dataset.

7.3. 3. Data Preprocessing

Text is standardized through removal of stop words, tokenization, lemmatization, and stemming, ensuring consistency across different languages.

7.3. 4. Feature Extraction using TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) highlights important terms specific to the dataset by balancing frequency and uniqueness.

7.3. 5. Feature Selection using PCA

Principal Component Analysis (PCA) reduces data dimensions while retaining key information, optimizing model performance and efficiency.

7.3. 6. Dataset Splitting

The dataset is divided into training and testing subsets to ensure balanced multilingual representation for model evaluation.

7.3. 7. Model Building

Three models—BERT, RoBERTa, and LSTM—are trained to analyze sentiments, utilizing deep learning and contextual comprehension for enhanced accuracy.

7.3. 8. Model Evaluation

Performance metrics like accuracy, precision, recall, F1-score, and confusion matrix are used to assess the effectiveness of the models.

7.3. 9. Sentiment Analysis

The models predict sentiment classes (positive, negative, neutral) in real time, demonstrating their capability for multilingual data processing.

7.3. 10. Continuous Monitoring

Model performance is tracked, and periodic updates are implemented to adapt to new data and ensure long-term efficiency in real-time analysis.

8. Expected Outcome

The research aims to develop an advanced framework for real-time sentiment analysis using multilingual image-text datasets by integrating cutting-edge models such as BERT, RoBERTa, and LSTM. Following systematic preprocessing steps like text extraction via OCR, stop-word removal, tokenization, and lemmatization, and leveraging feature extraction through TF-IDF with dimensionality reduction through PCA, the study is expected to achieve high classification accuracy, precision, recall, and F1-score across positive, neutral, and negative sentiment categories. This approach would achieve high real-time performance in sentiment detection and address class imbalance through the use of balanced multilingual

datasets for scalability with distributed systems. Therefore, a comprehensive methodology as such would provide a sound, scalable, and more inclusive solution for the analysis of multilingual sentiments by industries of marketing, public opinion monitoring or customer feedback analysis by improving adaptability and efficiency.

9. Research Work Plan

Phases	Plan of Work	Duration in Months					
		1-6	7-10	11-18	19-27	28-30	31-36
Phase I	Review of Literature	←	—	—	—	—	→
Phase II	Study of existing Methodology		←	→			
Phase III	Database creation/collection			←	→		
Phase IV	Design and Development of NLP and its model				←	→	
Phase V	Result and Conclusion					←	→
Phase VI	Report writing						← →

← → Arrow indicates the duration in months

10. References

1. Karageorgou, Ioanna N. "Just-in-time Sentiment Analysis for Multilingual Streams." (2020).
2. Devananthan, A. G., and Lekshmi S. Nair. "Exploring Multilingual Indian Twitter Sentiment Analysis: A Comparative Study." In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-8. IEEE, 2023.
3. Liu, Peng, Lemei Zhang, and Jon Atle Gulla. "Multilingual review-aware deep recommender system via aspect-based sentiment analysis." *ACM Transactions on Information Systems (TOIS)* 39, no. 2 (2021): 1-33.
4. Sharma, Amit, Neha Patel, and Rajesh Gupta. "Leveraging BERT and Sentiment Analysis Algorithms for Enhanced AI-Driven Brand Sentiment Monitoring." *European Advanced AI Journal* 11, no. 8 (2022).
5. Salur, Mehmet Umut, and Ilhan Aydin. "A novel hybrid deep learning model for sentiment classification." *IEEE Access* 8 (2020): 58080-58093.

6. Chanda, Supriya, Anshika Mishra, and Sukomal Pal. "Sentiment analysis of code-mixed Dravidian languages leveraging pretrained model and word-level language tag." *Natural Language Processing*: 1-23.

7. Rane, N. L., P. Desai, J. Rane, and S. K. Mallick. "Using artificial intelligence, machine learning, and deep learning for sentiment analysis in customer relationship management to improve customer experience, loyalty, and satisfaction." *Trustworthy Artificial Intelligence in Industry and Society* (2024): 233-261.

8. Yang, Yuyi, Nan Lin, Quinlan Batcheller, Qianzi Zhou, Jami Anderson, and Ruopeng An. "Sentiment Analysis of Tweets on Menu Labeling Regulations in the US." Nutrients 15, no. 19 (2023): 4269.

9. Al Zoubi, Ala'Mahmoud Mohammed. "Spam Reviews Detection Models in Multilingual Contexts applying Sentiment Analysis, Metaheuristics, and Advanced Word Embedding." (2024).

10. Yusuf, Aliyu, Aliza Sarlan, Kamaluddeen Usman Danyaro, Abdullahi Sani BA Rahman, and Mujaheed Abdullahi. "Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources." *IEEE Access* (2024).

11. Darshan, K., Jerusha Samuel, C. Manjunatha Swamy, Prashant Koparde, and N. Shivashankara. "NLP-Powered Sentiment Analysis on the Twitter." Saudi J Eng Technol 9, no. 1 (2024): 1-11.

12. Wong, Ethan Parker, and Faten M'hiri. "Analyzing Language Bias Between French and English in Conventional Multilingual Sentiment Analysis Models." *arXiv preprint arXiv:2405.06692* (2024).

13. Sharma, Neeraj Anand, ABM Shawkat Ali, and Muhammad Ashad Kabir. "A review of sentiment analysis: tasks, applications, and deep learning techniques." *International Journal of Data Science and Analytics* (2024): 1-38.

14. Aliyu, Yusuf, Aliza Sarlan, Kamaluddeen Usman Danyaro, and Abdulahi Sani Rahman. "Comparative Analysis of Transformer Models for Sentiment Analysis in Low-Resource Languages." *International Journal of Advanced Computer Science & Applications* 15, no. 4 (2024).

15. Das, Adway, Abhishek Kumar Prajapati, Pengxiang Zhang, Mukund Srinath, and Andisheh Ranjbari. "Leveraging Twitter Data for Sentiment Analysis of Transit User Feedback: An NLP Framework." *arXiv preprint arXiv:2310.07086* (2023).

16. Ahmet, Ahmed, and Tariq Abdullah. "Real-time social media analytics with deep transformer language models: A big data approach." In *2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE)*, pp. 41-48. IEEE, 2020.

17. Geethanjali, R., and A. Valarmathi. "A novel hybrid deep learning IChOA-CNN-LSTM model for modality-enriched and multilingual emotion recognition in social media." *Scientific reports* 14, no. 1 (2024): 22270.

18. Jha, Rahul Kumar. "Strengthening smart grid cybersecurity: An in-depth investigation into the fusion of machine learning and natural language processing." *Journal of Trends in Computer Science and Smart Technology* 5, no. 3 (2023): 284-301.

19. Liu, Jiabei, Keqin Li, Armando Zhu, Bo Hong, Peng Zhao, Shuying Dai, Changsong Wei, Wenqian Huang, and Honghua Su. "Application of Deep Learning-Based Natural Language Processing in Multilingual Sentiment Analysis." *Mediterranean Journal of Basic and Applied Sciences (MJBAS)* 8, no. 2 (2024): 243-260.

20. Hashmi, Ehtesham, Sule Yildirim Yayilgan, Ibrahim A. Hameed, Muhammad Mudassar Yamin, Mohib Ullah, and Mohamed Abomhara. "Enhancing multilingual hate speech detection: From language-specific insights to cross-linguistic integration." *IEEE Access* (2024).

21. Sampath, Koyyalagunta Krishna, and M. Supriya. "Transformer Based Sentiment Analysis on Code Mixed Data." *Procedia Computer Science* 233 (2024): 682-691.

22. Ghosh, Soumitra, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. "Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data." *Knowledge-Based Systems* 260 (2023): 110182.

23. Khan, Asif, Huaping Zhang, Nada Boudjellal, Arshad Ahmad, and Maqbool Khan. "Improving Sentiment Analysis in Election-Based Conversations on Twitter with ElecBERT Language Model." *Computers, Materials & Continua* 76, no. 3 (2023).

24. Yang, Lily Wei Yun, Wei Yan Ng, Xiaofeng Lei, Shaun Chern Yuan Tan, Zhaoran Wang, Ming Yan, Mohan Kashyap Pargi et al. "Development and testing of a multi-lingual Natural Language Processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: A multi-center study." *Frontiers in Public Health* 11 (2023): 1063466.

25. Kanfoud, Mohamed Raouf, and Abdelkrim Bouramoul. "SentiCode: A new paradigm for one-time training and global prediction in multilingual sentiment analysis." *Journal of Intelligent Information Systems* 59, no. 2 (2022): 501-522.

26. Li, Yuming, Johnny Chan, Gabrielle Peko, and David Sundaram. "Multilingual Sentiment Analysis in Social Media: catch the Emotion behind Different Expressions." In *PACIS*, p. 69. 2021.

27. El-Affendi, Mohammed A., Khawla Alrajhi, and Amir Hussain. "A novel deep learning-based multilevel parallel attention neural (MPAN) model for multidomain arabic sentiment analysis." *IEEE Access* 9 (2021): 7508-7518.

28. Song, Guizhe, and Degen Huang. "A sentiment-aware contextual model for real-time disaster prediction using twitter data." *Future Internet* 13, no. 7 (2021): 163.

29. Niu, Teng, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. "Sentiment analysis on multi-view social data." In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*, pp. 15-27. Springer International Publishing, 2016.

30. Müller, Martin, Marcel Salathé, and Per E. Kummervold. "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter." *Frontiers in artificial intelligence* 6 (2023): 1023281.

31. Özçift, Akın, Kamil Akarsu, Fatma Yumuk, and Cevhernur Söylemez. "Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish." *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije* 62, no. 2 (2021): 226-238.

32. Kalusivalingam, Aravind Kumar, Amit Sharma, Neha Patel, and Vikram Singh. "Enhancing Customer Relationship Management with Natural Language Processing: A Comparative Study of BERT and LSTM Algorithms." *International Journal of AI and ML* 1, no. 2 (2020).

33. Liu, Zejian, Gang Li, and Jian Cheng. "Hardware acceleration of fully quantized bert for efficient natural language processing." In *2021 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 513-516. IEEE, 2021.

34. Shahid, Mustafizur R., and Hervé Debar. "Cvss-bert: Explainable natural language processing to determine the severity of a computer security vulnerability from its description." In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1600-1607. IEEE, 2021.

35. Noorian, A., A. Harounabadi, and M. Hazratifard. "A sequential neural recommendation system exploiting BERT and LSTM on social media posts." *Complex & Intelligent Systems* 10, no. 1 (2024): 721-744.

36. Rothman, Denis. *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*. Packt Publishing Ltd, 2021.

37. Tarunesh, Ishan, Somak Aditya, and Monojit Choudhury. "Trusting roberta over bert: Insights from checklisting the natural language inference task." *arXiv preprint arXiv:2107.07229* (2021).

38. Angin, Merih, Beyza Taşdemir, Cenk Arda Yılmaz, Gökcan Demiralp, Mert Atay, Pelin Angin, and Gökhan Dikmener. "A roberta approach for automated processing of sustainability reports." *Sustainability* 14, no. 23 (2022): 16139.

39. Tan, Kian Long, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. "RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network." *IEEE Access* 10 (2022): 21517-21525.

40. Behera, Ranjan Kumar, Monalisa Jena, Santanu Kumar Rath, and Sanjay Misra. "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data." *Information Processing & Management* 58, no. 1 (2021): 102435.

41. Huang, Faliang, Xuelong Li, Changan Yuan, Shichao Zhang, Jilian Zhang, and Shaojie Qiao. "Attention-emotion-enhanced convolutional LSTM for sentiment analysis." *IEEE transactions on neural networks and learning systems* 33, no. 9 (2021): 4332-4345.

42. Murthy, G. S. N., Shanmukha Rao Allu, Bhargavi Andhavarapu, Mounika Bagadi, and Mounika Belusonti. "Text based sentiment analysis using LSTM." *Int. J. Eng. Res. Tech. Res* 9, no. 05 (2020).

43. Priyadarshini, Ishaani, and Chase Cotton. "A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis." *The Journal of Supercomputing* 77, no. 12 (2021): 13911-13932.

44. Long, Jie, A. Q. M. Khaliq, and Khaled M. Furati. "Identification and prediction of time-varying parameters of COVID-19 model: a data-driven deep learning approach." *International journal of computer mathematics* 98, no. 8 (2021): 1617-1632.

45. Soufyane, Ayanouz, Boudhir Anouar Abdelhakim, and Mohamed Ben Ahmed. "An intelligent chatbot using NLP and TF-IDF algorithm for text understanding applied to the medical field." In *Emerging Trends in ICT for Sustainable Development: The Proceedings of NICE2020 International Conference*, pp. 3-10. Cham: Springer International Publishing, 2021.

46. Campillo-Ageitos, Elena, Hermenegildo Fabregat, Lourdes Araujo, and Juan Martinez-Romo. "NLP-UNED at eRisk 2021: self-harm early risk detection with TF-IDF and linguistic features." *Working notes of CLEF* (2021): 21-24.

47. Gothankar, Ajinkya, Lavish Gupta, Niharika Bisht, Samiksha Nehe, and P. M. Bansode. "Extractive Text and Video Summarization using TF-IDF Algorithm." *International Journal for Research in Applied Science Engineering Technology (IJRASET)* 10, no. 3 (2022): 927-932.

48. Xu, Qianwen, Victor Chang, and Ching-Hsien Hsu. "Event study and principal component analysis based on sentiment analysis–a combined methodology to study the stock market with an empirical study." *Information Systems Frontiers* 22, no. 5 (2020): 1021-1037.

49. Naeem, Samreen, Wali Khan Mashwani, Aqib Ali, M. Irfan Uddin, Marwan Mahmoud, Farrukh Jamal, and Christophe Chesneau. "Machine learning-based USD/PKR exchange rate forecasting using sentiment analysis of Twitter data." *Computers, Materials & Continua* 67, no. 3 (2021): 3451-3461.

50. Choudhary, Sneha, and Charu Chhabra. "Sentiment analysis of Amazon food review data." In 2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT), pp. 116-120. IEEE, 2021.

51. Hesami, Mohsen, and Andrew Maxwell Phineas Jones. "Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture." Applied Microbiology and Biotechnology 104, no. 22 (2020): 9449-9485.

52. Kumar, Palvadi Srinivas, and Krishna Prasad. "Integrating OCR and NLP Techniques for Accurate Text Extraction and Plagiarism Detection in Image-Based Content." Library Progress International 44, no. 3 (2024): 2986-2996.